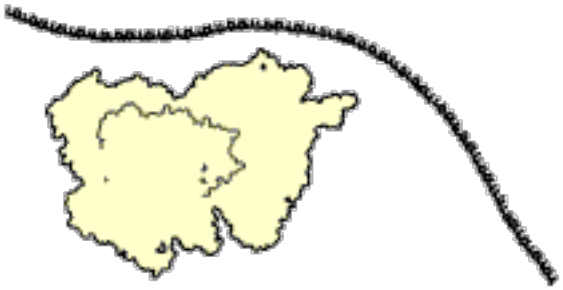
Causes and Effects of N-Terminal Codon Bias in Bacterial Genes

Mikk Eelmets

Journal Club

21.02.2014

Introduction



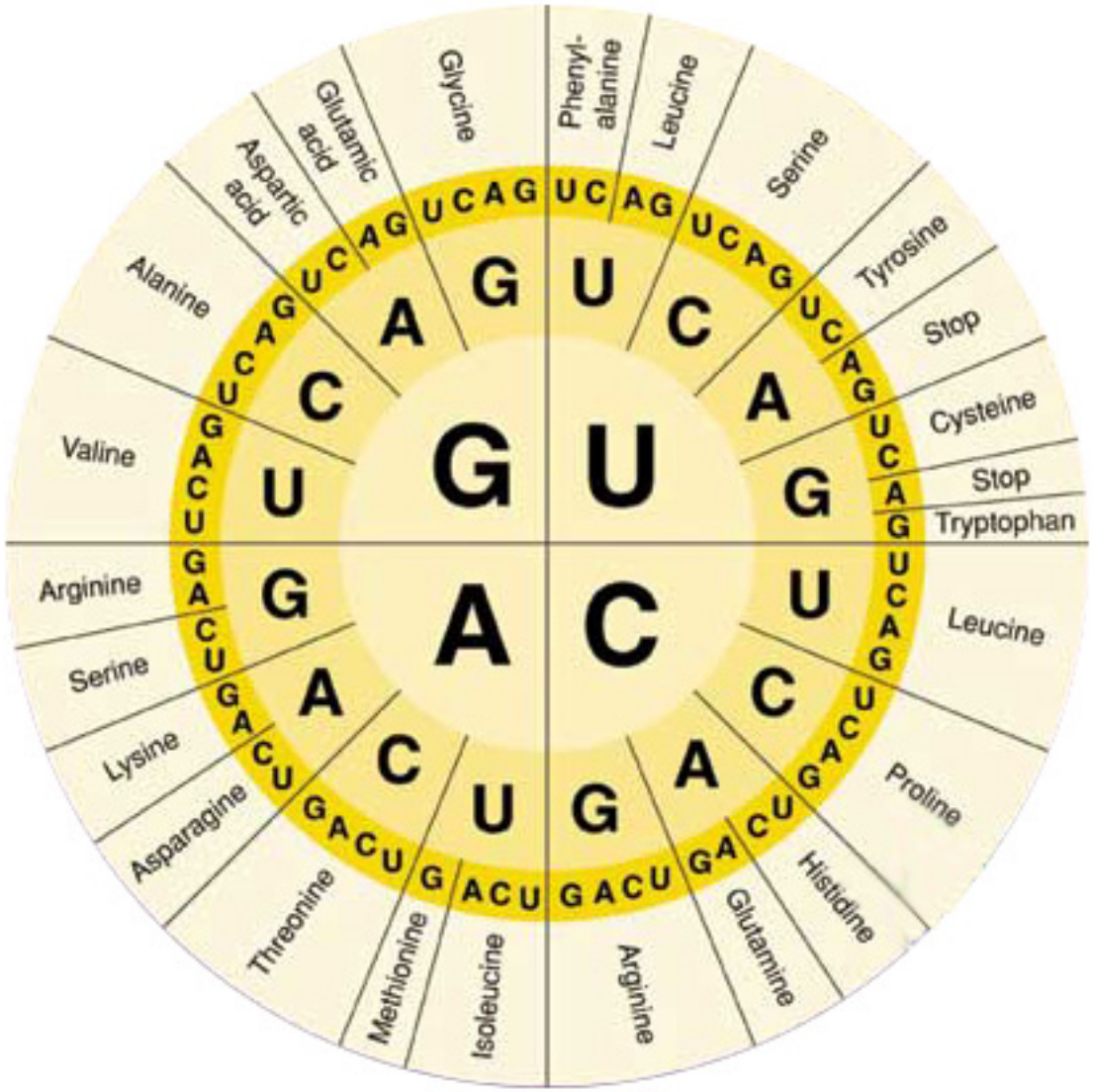
- Ribosomes were first observed in the mid-1950s (Nobel Prize in 1974)
- Nobel Prize in 2009 for determining the detailed structure and mechanism of the ribosome

There are still many questions without an answer

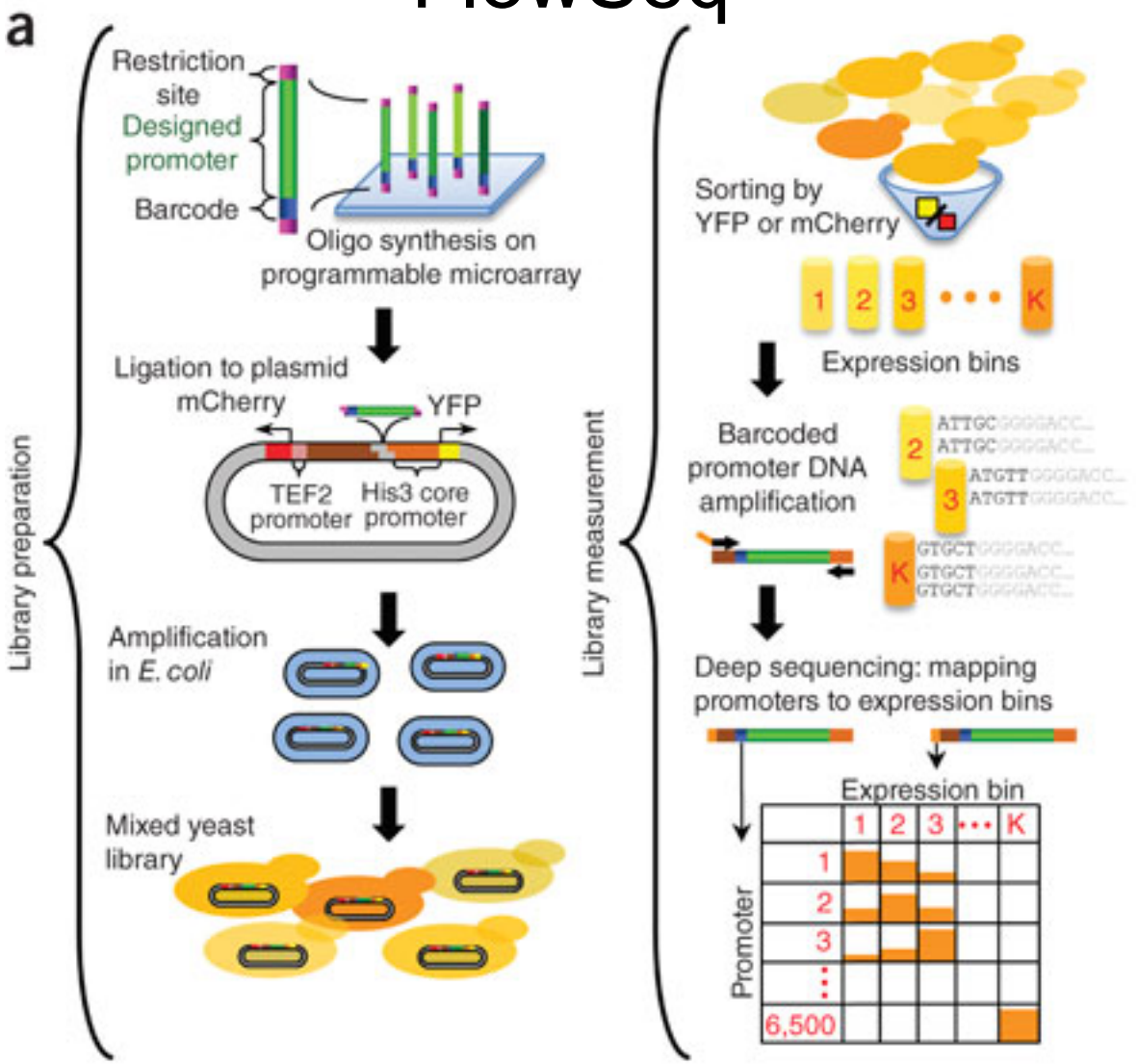
- Many organism are enriched for poorly adapted codon at the N terminus. Why?
- Which mechanisms are causal for expression changes?

Strategies: study genome-wide sequence correlation,
study conservation among species,
use synthetic genes

Genetic code

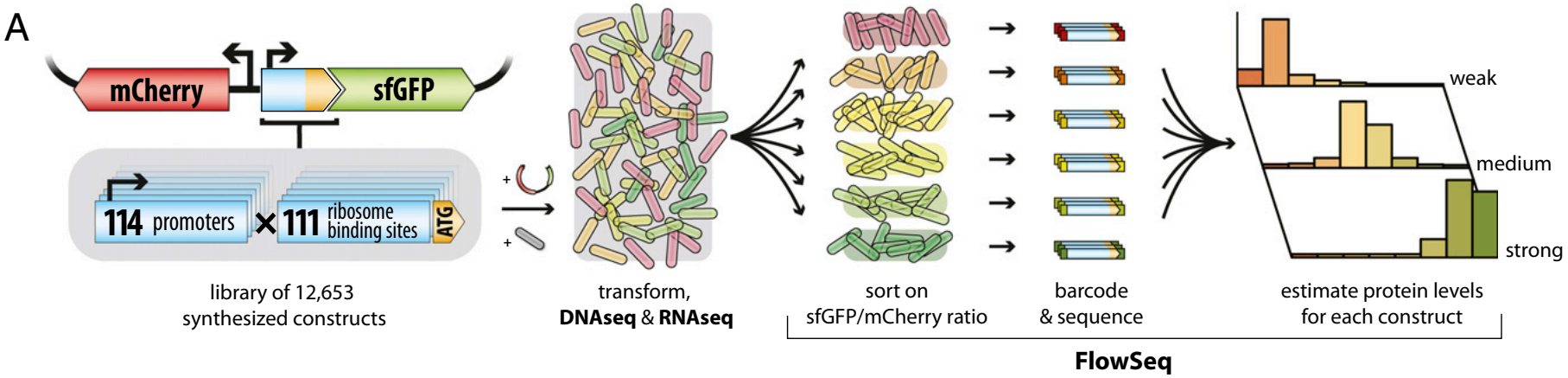


Design of experiment
FlowSeq



Sharon E et al Nat Biotechnol. 2012 May 20;30(6):521-30

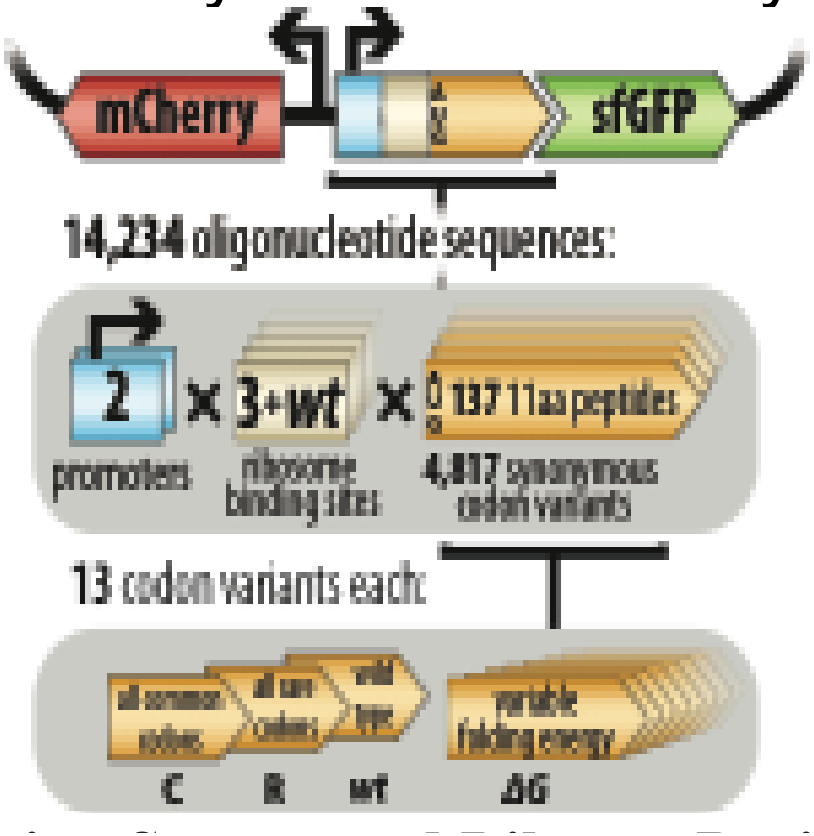
Design of experiment
FlowSeq



Kosuri S et al. Proc Natl Acad Sci U S A. 2013 Aug20;110(34):14024-9

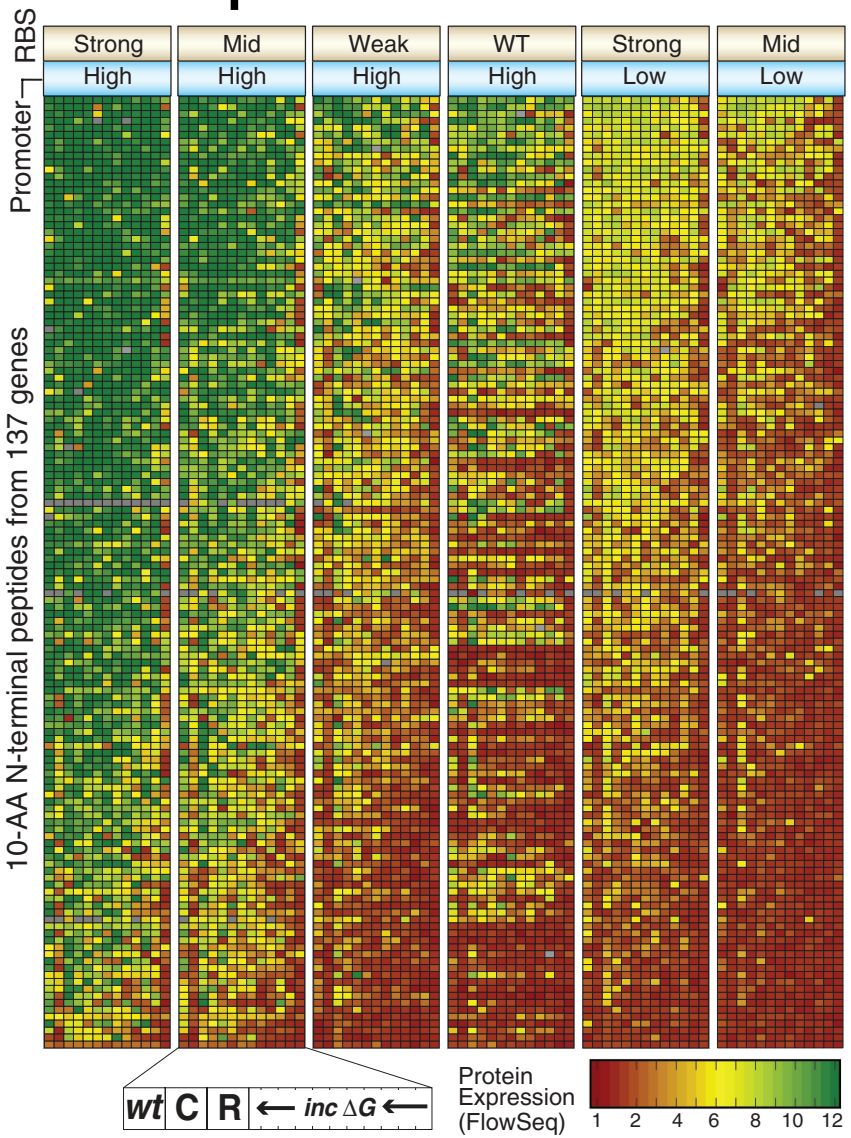
Design of experiment

Expression System and Library Design



Each library construct contained a **promoter**, **RBS**, and **11** codon N-terminal peptide (including the initiating 'ATG'). The peptide sequences correspond to the N terminus of **137** natural *E. coli* genes. For each promoter/RBS/peptide combination, we encoded **13** codon variants including the most common codons (**C**), most rare codons (**R**), wild-type sequence (**wt**), and codon variants with variable secondary structures (**ΔG**). The library was cloned in-frame with superfolder GFP (**sfGFP**). The GFP expression level is compared via relative fluorescence to a constitutively co-expressed **mCherry** protein.

Protein expression of the reporter library

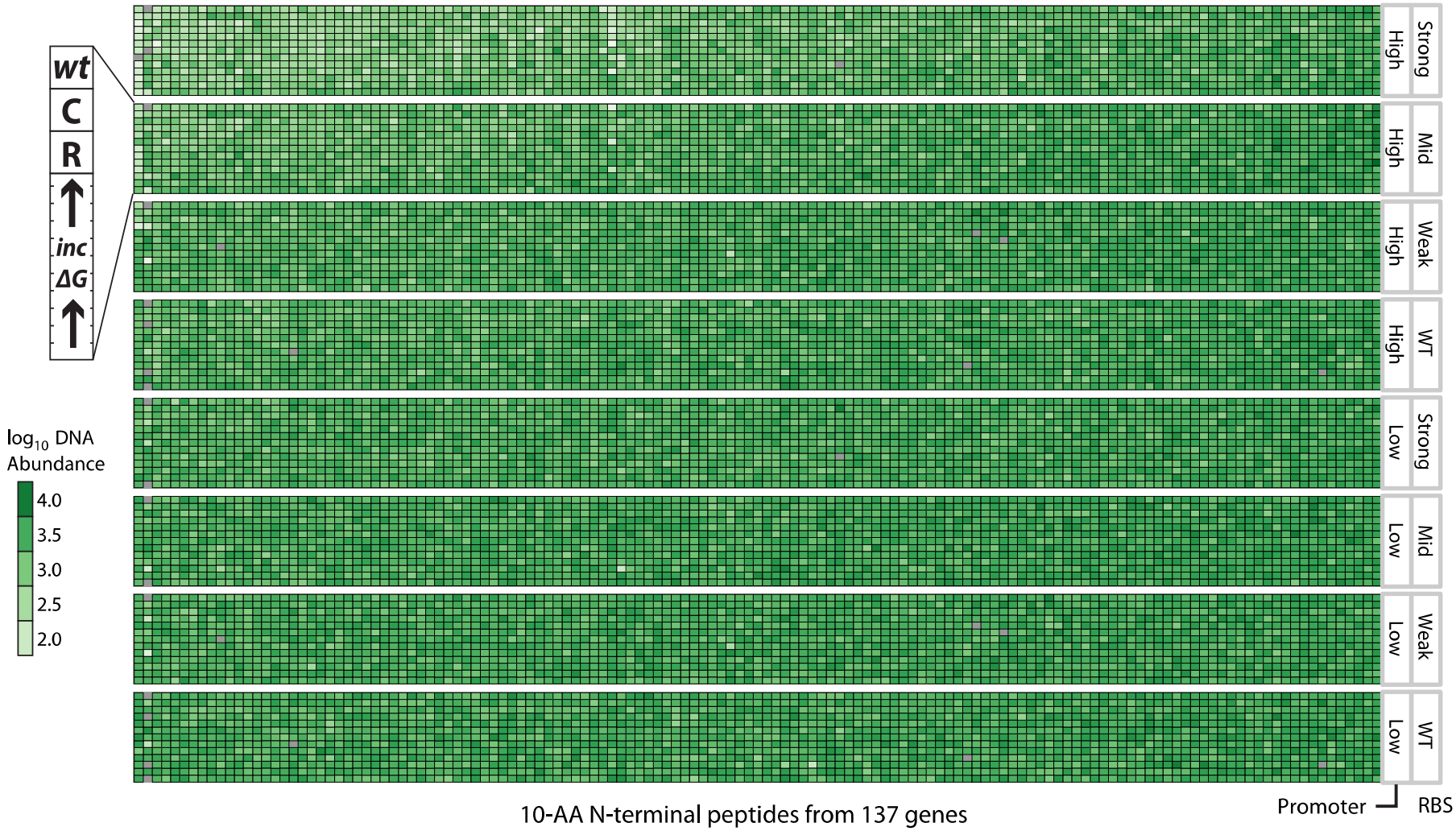


Protein levels:
7327 (51.5%) constructs

They used UNAFOLD software to predict free energy of folding for each RBS-codon variant

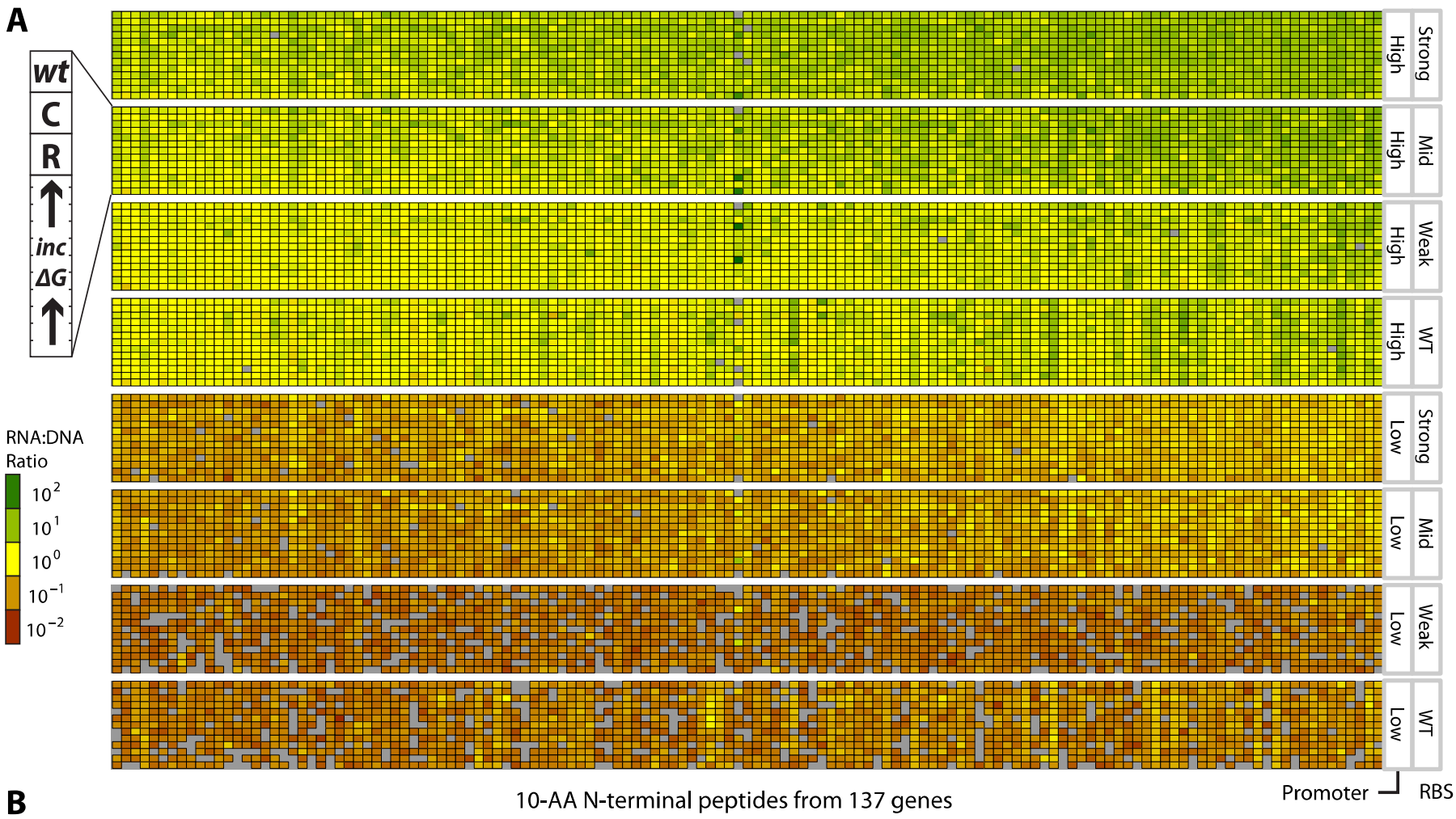
Protein expression of the library (as measured by the sfGFP:mCherry ratio) covers a ~200-fold range. The 13-member codon variant sets are grouped into columns by promoter/RBS combination (top). Codon variants include C, R, wild-type sequence (wt), and 10 sequences with varying secondary structure (ΔG).

RNA and DNA Abundance



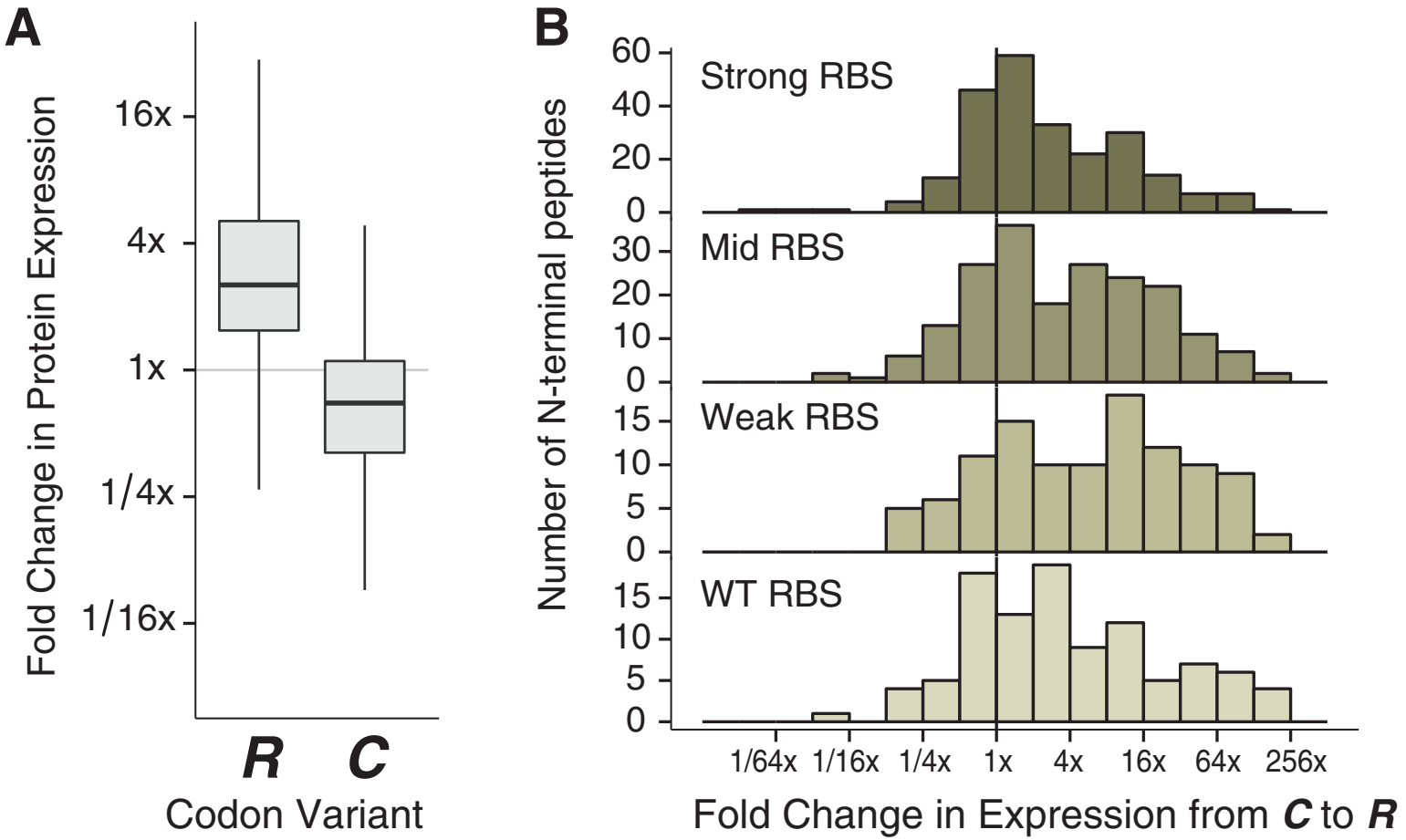
DNA abundances for each construct. Promoter identity is the first column and RBS identity is the second column. DNA abundances varied due to differences in DNA synthesis efficiencies as well as lower growth rate for very highly expressed genes

RNA and DNA Abundance



Relative RNA abundances (ratio of RNA to DNA contig counts) for each construct are displayed grouped by promoter and RBS identities.

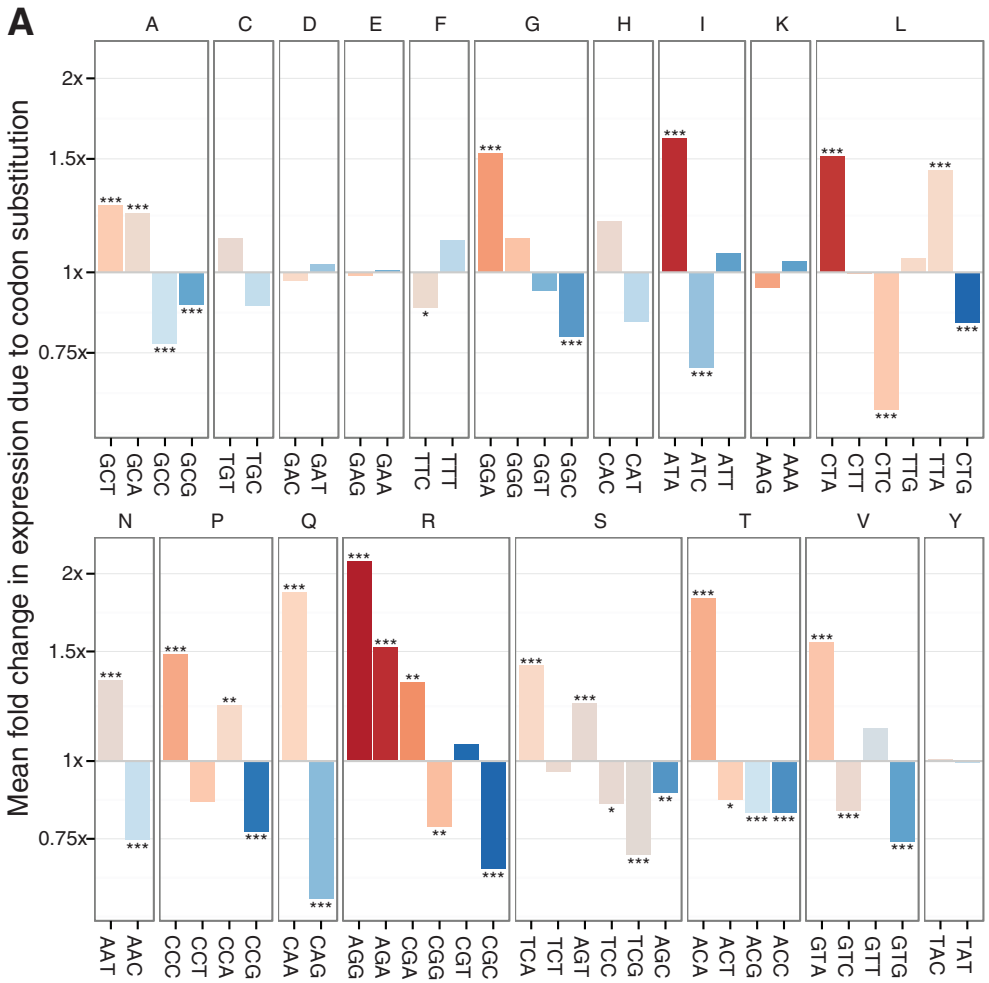
Gene expression measurements of the reporter library

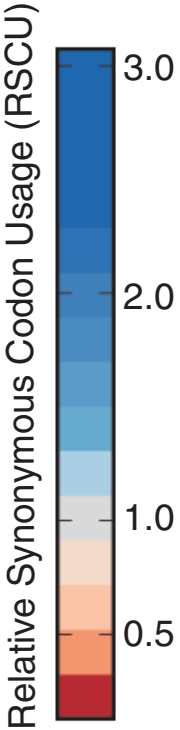


(A) N-terminal peptide sequences encoding the most rare (R) codon variants show increased expression when compared to the most common ones (C) (mean 14-fold; median 4-fold).

(B) (B) Fold change in expression between C and R codon variants is largely independent of RBS strength. WT, wild type.

Choice of codon and Expression





Relative Synonymous Codon Usage

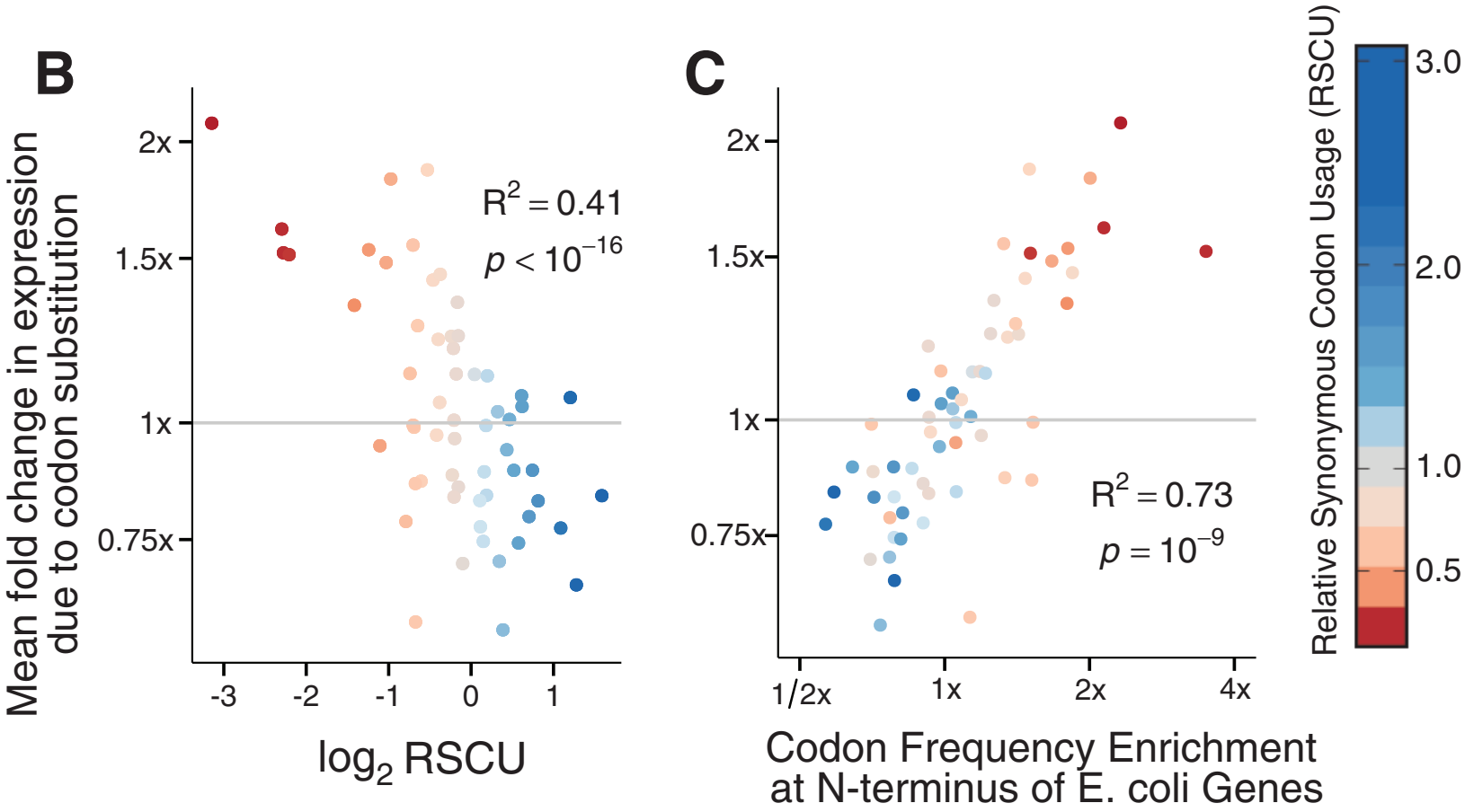
n = number of synonymous codons ($1 \leq n \leq 6$) for the amino acid under study, X_i = number of occurrences of codon *i*.

If the synonymous codons of an amino acid are used with equal frequencies, their *RSCU* values will equal 1.

$$RSCU_i = \frac{X_i}{\frac{1}{n}\sum_{i=1}^{n} X_i}$$

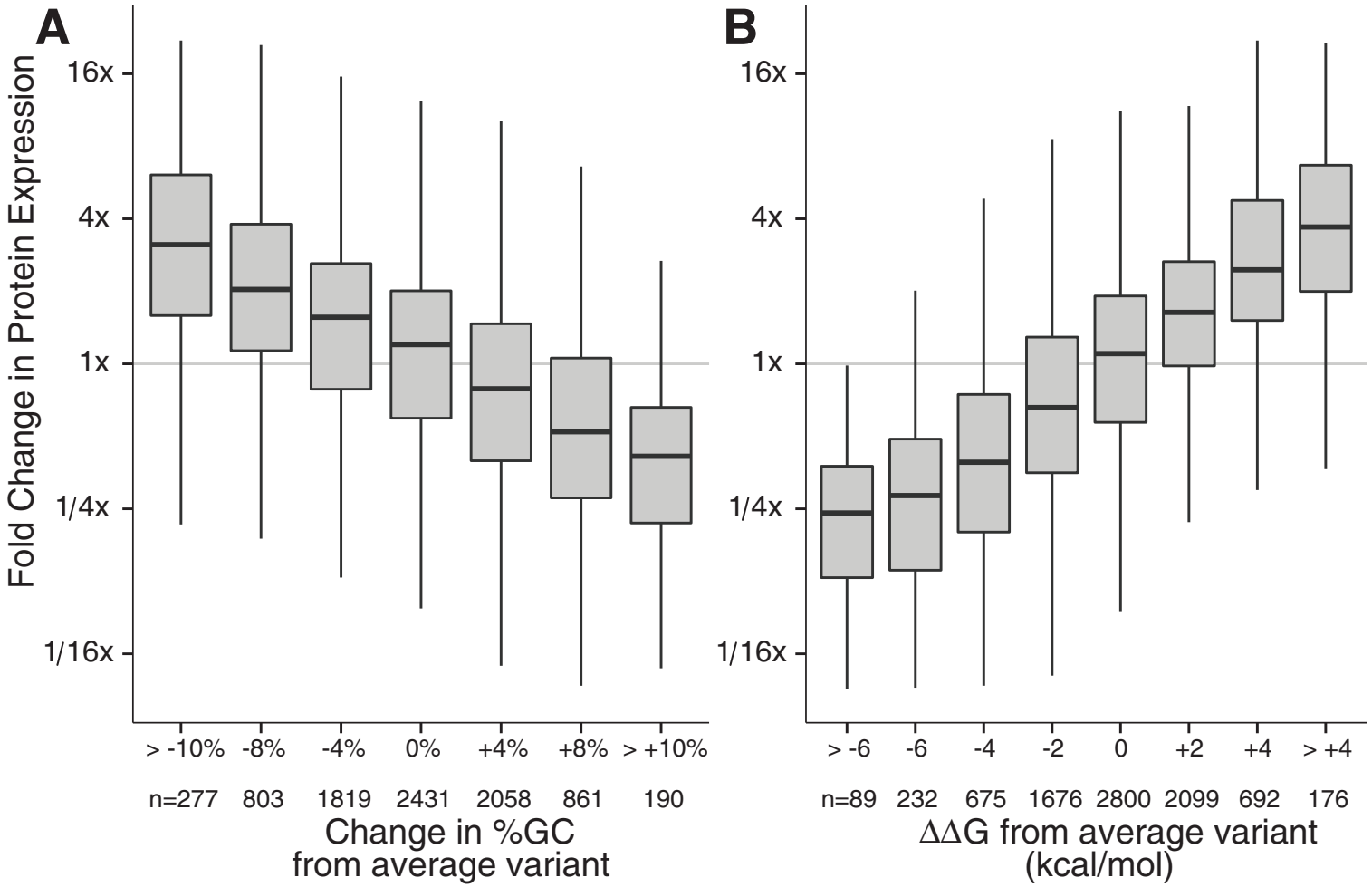
The average fold change in expression is correlated with the choice of codon. The y axis is the slope of a linear model linking codon use to expression change. Codons are sorted left to right by increasing genomic frequency and colored according to their relative synonymous codon usage (RSCU) in E. coli. (P values after Bonferroni correction: *P < 0.05, **P < 0.005, ***P < 0.001).

Choice of codon and Expression



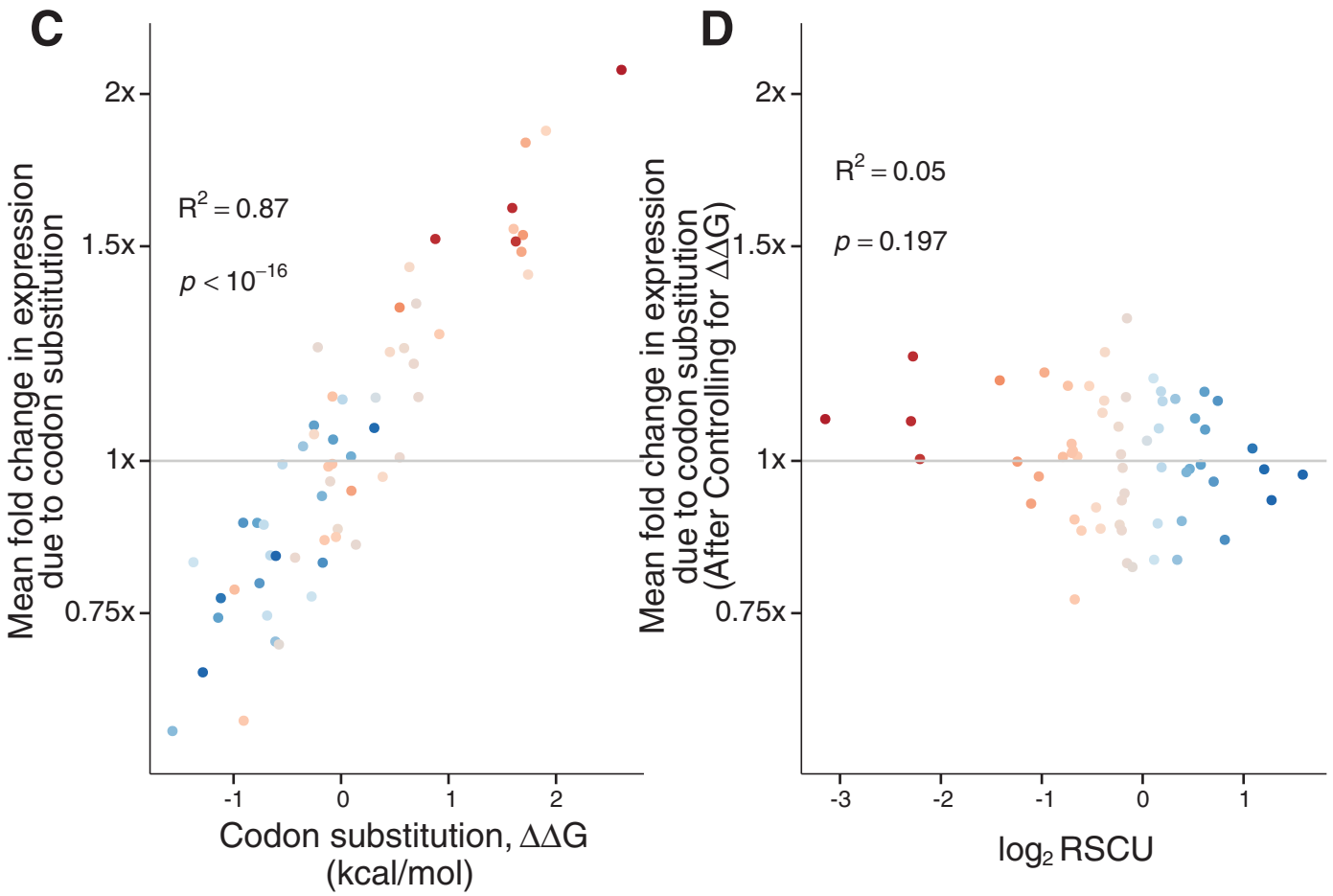
(B) The individual codon slopes (y axis) as in (A) show an inverse relationship with RSCU (x axis).
(C) The individual codon slopes correlate with enrichment of codons at the N terminus of genes in E. coli.

Choice of codon and Expression



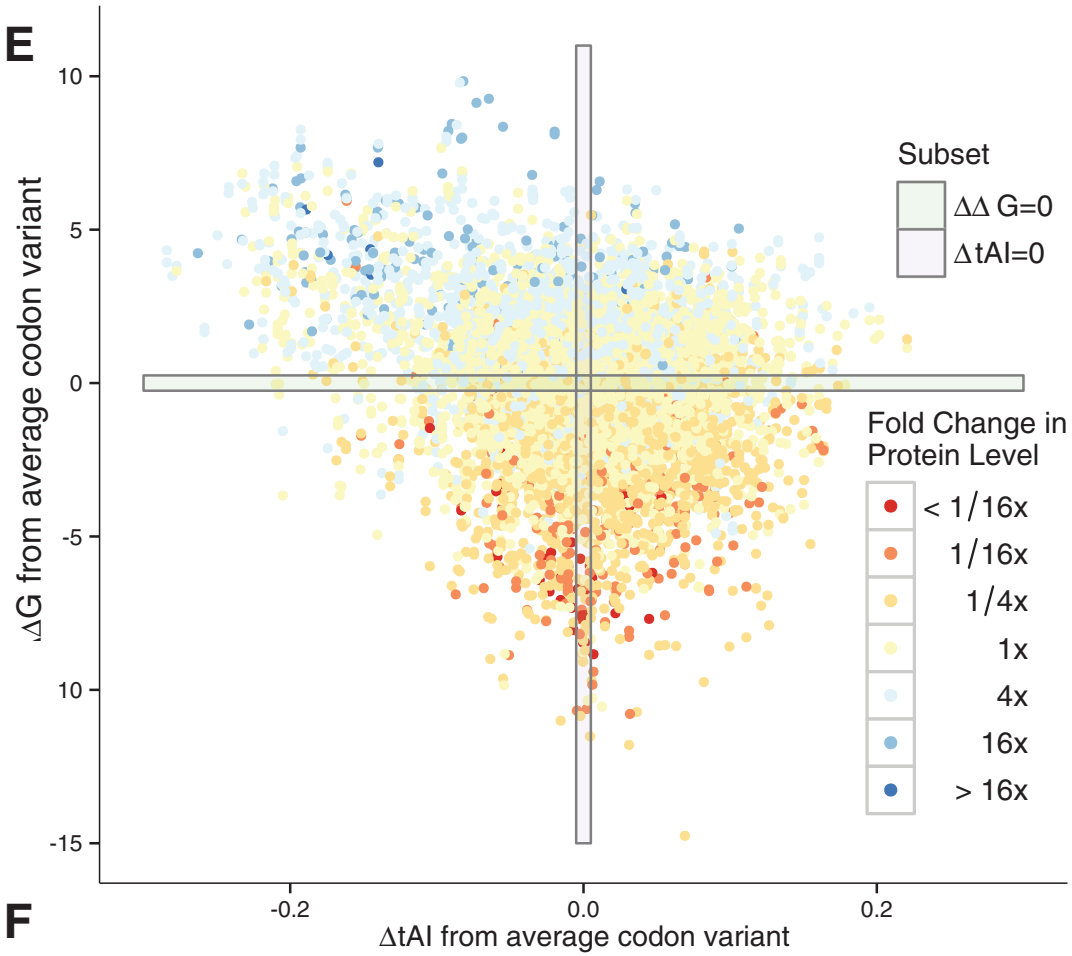
Rare codons alter expression by reducing mRNA secondary structure. (**A**) Expression changes are correlated with relative changes in %GC content. Each boxplot includes ±2% of centered value. (**B**) Expression increases correlate to relative increases in free energy of folding at the front of the transcript (ΔΔG). Each boxplot includes ±2 kcal/mol of centered value.

Choice of codon and Expression



Rare codons alter expression by reducing mRNA secondary structure. (**C**) Individual codon slopes correlate with the ΔΔG per individual codon substitution. (**D**) After controlling for ΔΔG with a multiple linear regression, there is no longer any relation between individual codon slopes and RSCU

Choice of codon and Expression



The classical translational efficiency ***cTE***$_i$ for each codon is the tRNA adaptation index (**tAI**)

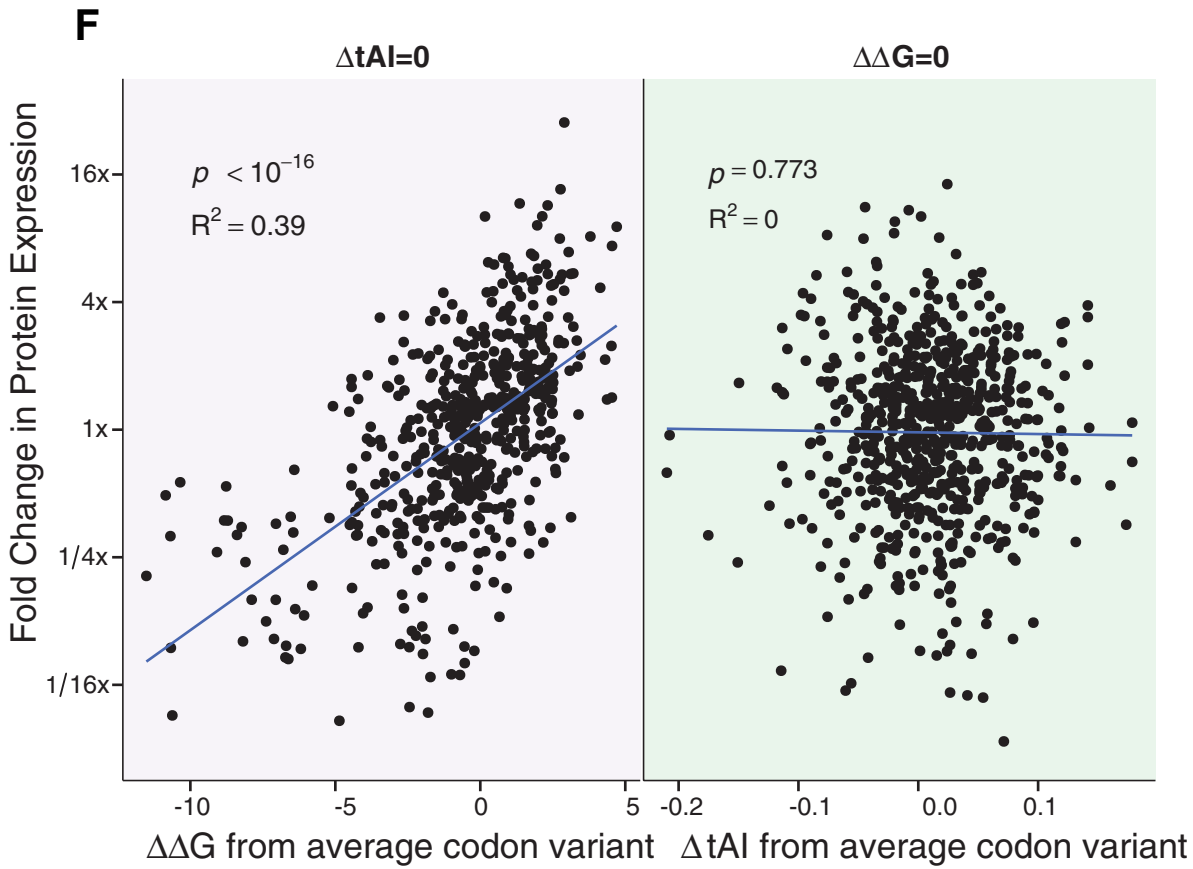
$$W_i = \sum_{j=1}^{n_i} (1 - s_{ij}) tGCN_{ij}$$

$$cTE = W_i / W_{\max}$$

W$_i$ –absolute adaptiveness
n$_i$ - the number of tRNA isoacceptors recognizing codon *i*
s$_{ij}$ - a selective constraint on the efficiency of the codon-anticodon coupling
tGCN$_{ij}$ – gene copy number of the tRNA

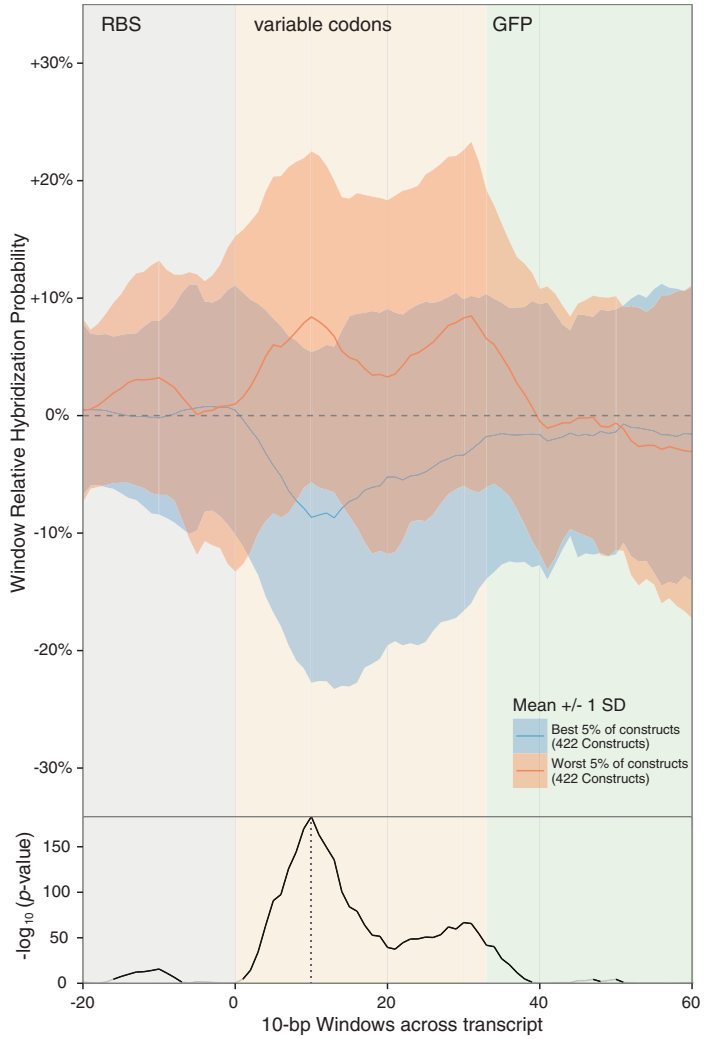
(**E**) The ΔΔG versus change in tAI is plotted for all constructs within the quantitative range. Constructs are colored by their relative fold change in expression from the average codon variant within the set.

Choice of codon and Expression



(**F**) Subsets of constructs corresponding to the shaded boxes in (E). (Left) Points with constant codon adaptation and varied secondary structure, (right) points with constant secondary structure and varied codon adaptation.

Hybridization probabilities and Expression



A multiple linear regression model that combines promoter and RBS choice, as well as N-terminal secondary structure and GC content, still explains only **54%** of variation in expression levels

Hybridization probabilities was calculated using UNAFOLD software

mRNA structure downstream of start codon is most correlated with reduced expression. (**Top**) The best and worst 5% of constructs as ranked by relative expression within a codon variant set are grouped and plotted as blue and red ribbons, respectively. The ribbon tops and bottoms are one standard de- viation from the mean, which is shown as a solid line. (**Bottom**) The p-value for linear regressions, correlating hybridization probabilities within each window to expression fold change in all constructs.

Conclusion

- Using N-terminal rare codons instead of common ones increase expression by ~14 fold (median 4 fold)
- Reduced RNA structure, not codon rarity itself, is responsible for expression increases
- This knowledge helps better predict how to synthesize genes that make enzymes or drugs in a bacterial cell.

Reference

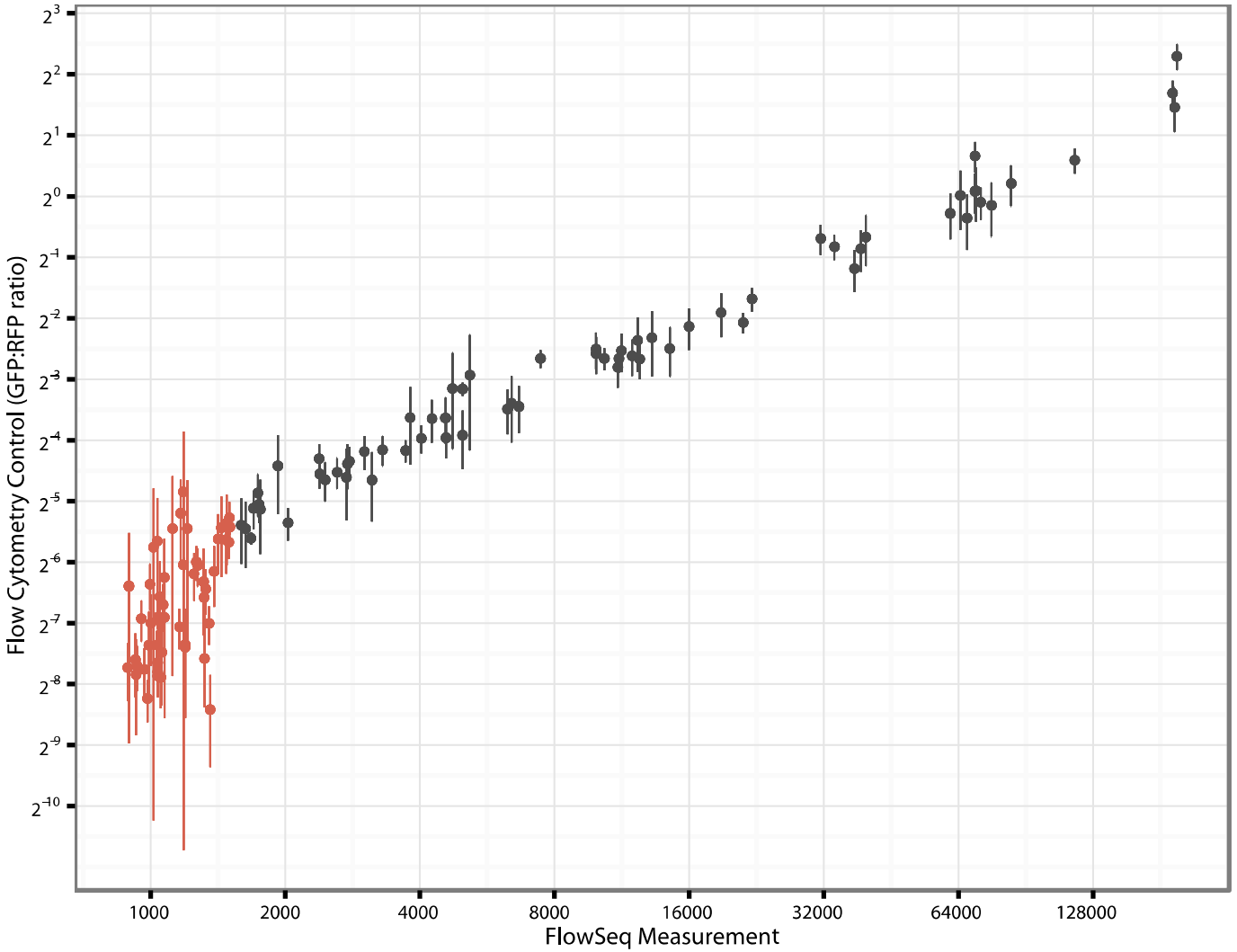
Daniel B. Goodman, George M. Church, Sriram Kosuri

“Causes and Effects of N-Terminal Codon Bias in Bacterial Genes“

Science. 2013 Oct 25;342(6157):475-9

THANK YOU

Flow Cytometry Controls



FlowSeq estimates of fluorescence ratio correlate well (R2 = 0.955, $p < 2\times10^{-16}$) with individually measured fluorescence ratios from sequence-verified clones. 51 constructs that were outside of the quantitative FlowSeq range are shown in red.